Pronóstico del aumento de ingresos en redes hospitalarias por diagnósticos de enfermedades respiratorias según las condiciones meteorológicas y la calidad del aire del Valle de Aburrá

Kenneth David Leonel Triana
Juan José Naranjo Velásquez

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Alejandro Ruiz Luna
Especialista en Analítica y Ciencia de Datos

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2025

Cita	(Leonel Triana & Naranjo Velásquez, 2024)
Referencia **Estilo APA 7 (2020)**	Leonel Triana, K. L., & Naranjo Velásquez, J. J. (2024). *Pronóstico del aumento de ingresos en redes hospitalarias por diagnósticos de enfermedades respiratorias según las condiciones meteorológicas y la calidad del aire del Valle de Aburrá***.** Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Grupo de Investigación Intelligent Information Systems Lab In2Lab

Especialización en Analítica y Ciencia de Datos, Cohorte VIII.

Centro de Investigación Ambientales y de Ingeniería (CIA)**.**

Dedicatoria

Le dedicamos este trabajo primeramente a Dios, por guiarnos y darnos fortaleza en cada etapa del camino. A nuestras familias y amigos, por su apoyo incondicional, comprensión y motivación constante. Este logro también va dedicado a todos aquellos que creyeron en nosotros incluso en los momentos más difíciles, siendo parte fundamental de esta etapa culminada.

Agradecimientos

Agradecemos profundamente a nuestro asesor de tesis, por su orientación, paciencia y compromiso durante el desarrollo de este proyecto. También, extendemos nuestro agradecimiento a los profesores que nos formaron a lo largo de la especialización, así como a nuestros compañeros de estudio, con quienes compartimos no solo conocimientos, sino también experiencias y aprendizajes que marcaron nuestra formación profesional. Este trabajo es reflejo del esfuerzo como equipo, de jornadas exigentes, y del compromiso por alcanzar nuestras metas.

Tabla de contenido

Lista de Figuras

Lista de Tablas

Siglas, acrónimos y abreviaturas

SIATA	Sistema de alerta temprana de Medellín y del Valle de Aburrá
RIPS	Sistema de Información de Prestaciones de Salud
EPOC	Enfermedad pulmonar obstructiva crónica
ETL	Proceso de la gestión de datos (recopilar, transformar y cargar)
ML	Machine Learning
MSE	Error cuadrático medio
SVM	Máquinas de vectores de soporte
INS	Instituto Nacional de Salud
XGBoost	Extreme Gradient Boosting
R2	Coeficiente de determinación
EDA	Análisis Exploratorio de Datos
IPS	Institución Prestadora de Salud
RMSE	Raíz cuadrada del error cuadrático medio
CPU	Unidad Central de Procesamiento
PCA	Análisis de componentes principales

Pronóstico de la capacidad de ingresos en redes hospitalarias basado en diagnósticos de enfermedades respiratorias, condiciones temporales, locativas y meteorológicas del Valle de Aburrá

***Resumen*— Colombia es un país que presenta un alto porcentaje de mortalidad por enfermedades respiratorias debido a sus zonas de concentración de contaminantes. Estas cifras pueden darse en ocasiones donde las condiciones climáticas y ambientales son dañinas para la salud. Dadas estas cifras alarmantes, es necesario tener en cuenta la cantidad de pacientes que pueden ingresar a los hospitales en ocasiones donde existan afectaciones en las condiciones ambientales, por ello, con el fin de alertar a los establecimientos de salud acerca de la cantidad de pacientes con diagnósticos asociados a enfermedades respiratorias que ingresarán en ciertos sectores o comunas de Medellín, presentamos un modelo alimentado por variables climáticas e ingresos en hospitales, con datos proporcionados por el Sistema de Alerta Temprana de Medellín y el Valle de Aburrá y el Sistema de Información de Prestaciones de Salud en los años de 2020 a 2024. La creación del modelo predictivo es a base de la metodología CRISP-DM, lo que permite realizar el proceso ETL en los datos para ajustar el modelo de Random Forest a un R2 de 0.9940. De esta forma, podemos estimar la cantidad de asistencias proyectadas en entidades hospitalarias cercanas a los puntos de toma de información.**

Keywords*— *machine learning, hospitales, enfermedades respiratorias, calidad del aire, SIATA, RIPS, Medellín.

1. Introducción

La calidad del aire ha sido un tema de abundante estudio en el Valle del Aburrá, debido a la distribución geográfica del espacio y que al estar rodeado por montañas genera una estrechez que impide el flujo libre de gases y partículas en el aire (Área Metropolitana Valle del Aburrá, 2019), por ende, estas partículas se concentran en el centro de esta misma, consecuentemente, es un efecto de convivencia que debe atacarse desde diferentes medios. Es por ello que desde finales de los años 90 se creó el Sistema de Alertas Tempranas SIATA como un proyecto para la gestión ambiental y de riesgos del Área Metropolitana del Valle de Aburrá, realizando monitoreos en tiempo real con modelamientos hidrológicos y meteorológicos para pronosticar ocurrencias de fenómenos naturales que puedan generar riesgos a la población (Universidad de EAFIT, 2020).

Según la Organización Mundial de la Salud (OMS), "Colombia es el décimo primer país de la región que presenta mayor mortalidad por enfermedades respiratorias, por lo que representa la cuarta causa de muerte en el país", esta medida genera alertamiento, como en 2019 donde estas enfermedades causaron 534.242 defunciones, con una tasa de 35.8 defunciones por 100,000 habitantes, y en 2020 con 22.9 defunciones por cada 100,000 habitantes (Ministerio de Salud y Protección Social, 2022). Adicionalmente, según el Instituto Nacional de Salud (INS) se atribuyen 15,681 muertes asociadas a la mala calidad del aire, lo que radica en 8 mil muertes anuales, que bajo una vista financiera abarca costos de 12.2 billones de pesos, equivalente al 1.5% del PIB. Por ello, la subdirección de Salud ambiental busca formas de tener una gestión de la calidad del aire del país orientadas a la disminución de la exposición y los factores de riesgo (Minsalud, 2021).

Se debe agregar que estas enfermedades respiratorias son silenciosas y en algunos casos no tiene cura como la enfermedad pulmonar obstructiva crónica (EPOC) (BBC News Mundo, 2017). En relación con lo anterior, bajo nuestra opinión, la visualización del ser humano no permite detectar en qué ocasiones hay una mala calidad del aire, por ello es posible inhalar componentes peligrosos para los pulmones que contribuyan a desarrollar alguna de estas enfermedades, y poco a poco, de manera discreta, se desarrolle estas sin conocimiento de aparición en el cuerpo, lo que hace que sea tarde para su prevención y tratamiento.

El aire que respiramos se compone en su mayor proporción por nitrógeno y oxígeno (Torres Jaramillo, 2021), sin embargo, según la zona geográfica y las condiciones climáticas, esta proporción puede verse afectada con partículas inhalables finas que causan efectos negativos en la función pulmonar (Agencia de Protección Ambiental de Estados Unidos, 2024), por ello, desde nuestro punto de vista, es un fenómeno que no podemos evitar (dada la distribución territorial de Medellín), pero que se puede monitorear y medir su impacto para anticiparse a complicaciones posteriores, así como en recientes estudios donde se realizan estimaciones de la calidad del aire bajo modelos de redes neuronales (2019, Marzo 25), lo que permite actuar con antelación en zonas de altos niveles de contaminantes.

De acuerdo con lo anterior, se plantea generar un modelo que estime posibles escenarios en los que se aumenten las capacidades de atención a pacientes de los sectores o comunas de Medellín por diagnósticos de enfermedades respiratorias respecto a las condiciones meteorológicas y ambientales del Valle del Aburrá, con el fin de generar pronósticos prescriptivos y de alertamiento para la preparación de estos establecimientos ante situaciones de alta demanda por malas condiciones climáticas. Este planteamiento se propone bajo una estructuración y modelación de datos en la metodología CRISP-DM a base de información climática, ambiental y registros médicos, cuya fuente de información es el Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA) y el Sistema de Información de Prestaciones de Salud (RIPS), donde concatenados, relacionan variables meteorológicas como temperatura, presión, humedad, vientos, entre otras, junto a registros de pacientes ingresados en hospitales por diagnósticos asociados a deficiencias respiratorias. Bajo este conjunto de datos, se toma un modelo que obedece a una técnica de regresión numérica bajo el algoritmo supervisado de machine learning XGBoost (extreme Gradient Boosting) y su evaluación en precisión se verifica bajo el R2 y el error cuadrático medio (MSE). De esta forma, se procura que los hospitales tengan la información de estos estimados para estar debidamente preparados ante un evento atípico de sobrecapacidad.

Bajo esta intención, en este artículo se encuentra información relacionada con los métodos de trabajo, adquisición y orígenes de datos, análisis exploratorio, tratamiento de variables, preprocesamiento de datos, selección de variables, resultados y comparativos de modelos.

2. Materiales y Métodos

El desarrollo de este artículo se apoya bajo herramientas tecnológicas como Python y librerías para manipulación y modelamiento de datos como Pandas, Scikit-learn, Seaborn, Scipy, XGBoost, Lightgbm, Shap, entre otras; también, se elabora este trabajo bajo la metodología CRISP-DM, una técnica reconocida ante la extracción de información valiosa a partir de volúmenes de datos, basándose en las etapas de entendimiento de negocio, entendimiento de datos, preparación de los datos, modelamiento y evaluación o validación desde el negocio para posterior despliegue del modelo. Esta forma de trabajo se toma con el fin de mitigar los inconvenientes que tengamos frente al desarrollo analítico y poder concluir los objetivos planteados ante la acción de las circunstancias encontradas.

a. Descripción de los datos

Los datos utilizados en el estudio se obtuvieron mediante fuentes de datos públicas como el portal de descargas del Sistema de Alerta Temprana de Medellín y el Valle de Aburrá SIATA junto con información de ingresos en hospitales del RIPS Medellín en el lapso de tiempo del 2020 al 2024, donde se seleccionaron datos meteorológicos, ambientales y médicos. Esta propuesta de solución inicia desde el entendimiento de la toma de datos con los puntos de monitoreo (entre automáticos y manuales) que usan la información de ceilómetros y radiómetros de la entidad SIATA, que se encuentran distribuidos a lo largo de todas las zonas del valle de Aburrá, estos son importantes porque son los encargados de tomar los datos meteorológicos y de la calidad del aire en diferentes medidas de control, por ello es una información que se muestra en variables numéricas y con un tamaño de 1493 registros al concatenar sus campos para una granularidad por día. A su vez, en esta solución se implementa la información del número de atenciones totales de pacientes por enfermedades relacionadas con los problemas respiratorios, los cuales tienen 3’436.000 registros de información encriptada para el cuidado de la

información de los pacientes que han entrado a consulta por estos motivos, y por ende maneja información de carácter numérico y categórico al tratar con variables como diagnósticos, edades, fechas de ingreso, entre otras.

b. Preprocesamiento y limpieza de datos

Para una visión más profunda de los datos que se encuentran se realiza un análisis exploratorio de datos (EDA) iniciando con la aclaración de que en el dataset concatenado de las variables meteorológicas y ambientales encontramos información relacionada con la presión, la humedad, la temperatura, partículas de tamaño menor a 2.5 micras (pm25), información relacionada con los vientos y la cantidad de agua por lluvia desde los pluviómetros, sin embargo, algunas de las 3 estaciones de recolección de datos (68-Jardín Botánico, 197-Universidad de Medellín y 198-Politecnico Jaime Isaza Cadavid) presentan datos faltantes por posibles mantenimientos o daños en los equipos en ciertas épocas del periodo de 2020 a 2024. Aun así, podemos observar el comportamiento de las variables en estos conjuntos de datos como se muestra en la Figura 1.

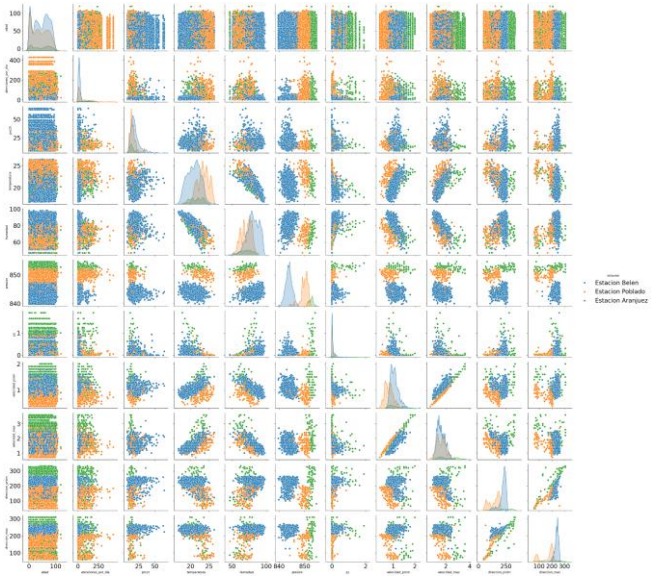

Figura 1 Combinaciones de variables y diagramas de distribución.

Adicionalmente, desde la fuente de datos se proporciona una columna de calidad de datos que indica si el dato se toma en tiempo real o con un desfase, por ende, se trabaja con el código 1 (calidad confiable del dato en tiempo real) para un análisis sin ruido ni sesgos. Por otro lado, la fuente de datos del RIPS no presenta información faltante debido a los registros por pacientes encriptados, no obstante, se nota una mayor concentración de registros de pacientes en algunas Instituciones Prestadoras de Servicios de salud (IPS) que pueden causar sesgo.

En el transcurso de la construcción de la base, se generan dos columnas categóricas llamadas "dia_semana" y "festivo", esto con la intención de entregarle un peso al modelo donde observe la diferencia de condiciones climáticas cuando es inicio de semana, media semana y final de semana, además, de que permite asociar el festivo como un día semejante a los fines de semana, a partir de ello se podría asociar diferencias de contaminaciones cuando hay mayor flujo personal y vehicular en las zonas de Medellín dependientes de los días de circulación. Continuando con estas transformaciones, se realiza en los datos médicos una agrupación por conteo creando la variable "atenciones_por_dia" de tal forma que no se tenga una granularidad por paciente, sino que se note por la variable "nombreIPS" la cantidad de pacientes de la institución en ese día de las condiciones, adicionalmente, para tener una categorización más expansiva de la localización se crean las variables "comuna_ips" y "sector_ips", reflejando la comuna donde se encuentra la IPS y el sector geográfico (centrooriental, centrooccidental, suroccidental, suroriental, nororiental, noroccidental). Además, dado que se tiene registros médicos con visitas a IPS de pocos pacientes se pone un límite de mínimo 1020 pacientes para ingresar al modelo.

Respecto a los valores crudos encontrados en el dataset de las condiciones meteorológicas, se hallaron datos faltantes, para evitar entrar en conflictos con este tratamiento y aprovechando la cantidad de información disponible de registros por hora y minuto de estas condiciones, se decide eliminar estos registros y dejar únicamente aquellos que tuvieran datos adecuados bajo el filtro de la columna calidad. Por otro lado, dado que el dataset de condiciones ambientales se encontraba en una granularidad de horas y el dataframe de registros médicos por día, se realiza una agrupación de los datos por la media aritmética de los datos, de tal forma que ambos conservan el mismo grado de granularidad y la posibilidad de cruzar la información relacionado por fecha, estación de medición e IPS de atención. Seguidamente, al tener un conjunto global de datos

relacionados se lleva a cabo una identificación de datos atípicos mediante el modelo Isolation Forest de la librería Scikit-learn, mostrando que existen unos datos que se ubican en la cola izquierda muy alejados de la distribución como se muestra en la Figura 2, por ende, se eliminan datos asignados como outliers y se trabaja con el dataset sin valores nulos ni datos atípicos.

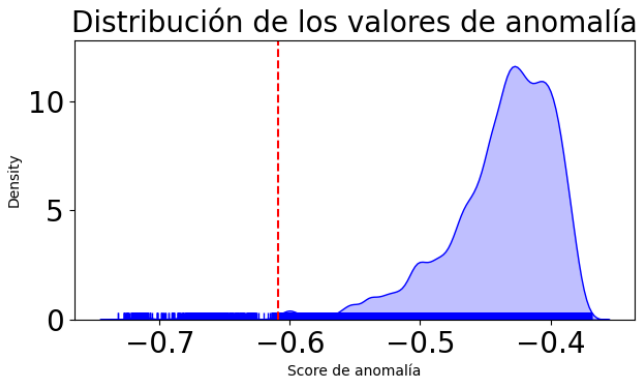

Figura 2 Distribución de score de anomalía.

Luego de la eliminación de datos anómalos, se busca valores duplicados y se encuentra que al realizar los cruces de datos hallamos registros duplicados en todas las columnas, por ende, eliminamos estos valores quedándonos con uno representativo. Ahora con el dataset de trabajo, revisamos las distribuciones de las variables categóricas y discretas. También, se debe tener en cuenta que el dataset se dispone con variables numéricas y categóricas, por ende, se procede a codificar las variables categóricas en numéricas sin órdenes de enumeración mediante la librería de scikit-learn aplicando el Ordinal Encoder.

Para las variables cualitativas realizamos los diagramas de distribución y box-plot para observar el comportamiento de estas detallando que se toma datos de una muestra aleatoria para observar si la toma de datos aleatorio se comporta semejante al patrón general de las variables, se pone en muestra para las variables atenciones_por_dia, humedad, presión y pm25 en la Figura 3.

Una vez se tienen estas variables recopiladas y seleccionadas, se realiza una matriz de correlación para ver el comportamiento binario entre ellas donde, así como se muestra en la Figura 4. Se puede inferir que hay variables que están muy relacionadas que es la 'velocidad_prom' con 'velocidad_max' esto nos ayuda tomar decisiones de quedarnos con solo una de esas características y comprender mejor el comportamiento de nuestra variable objetivo con las demás variables

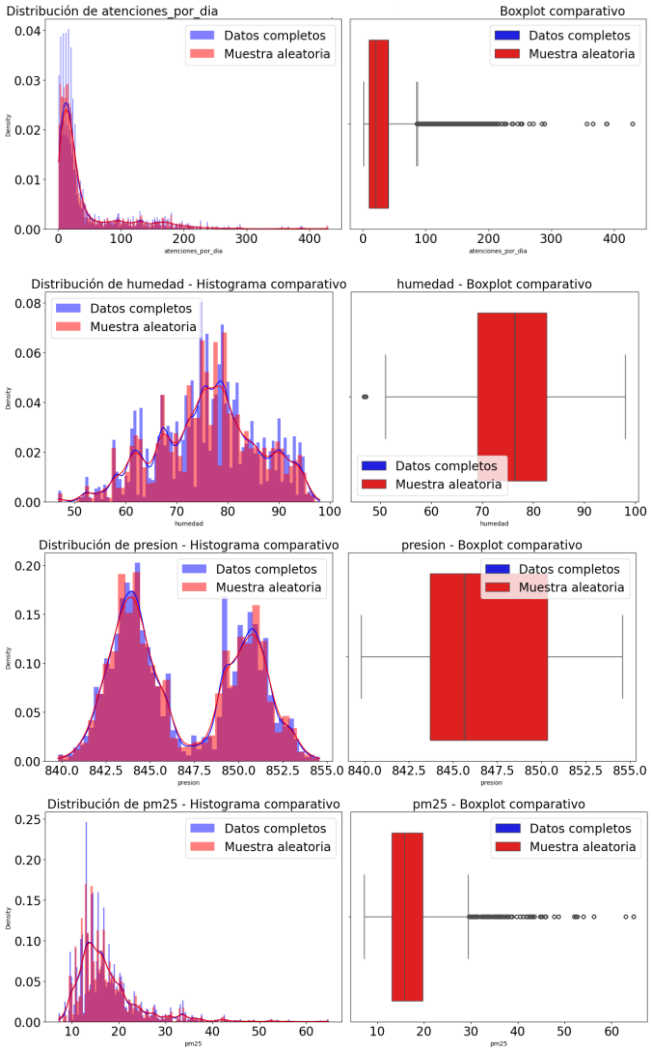

Figura 3 Diagramas de distribución y box-plot de variables numéricas atenciones_por_dia, humedad, presión y pm25.

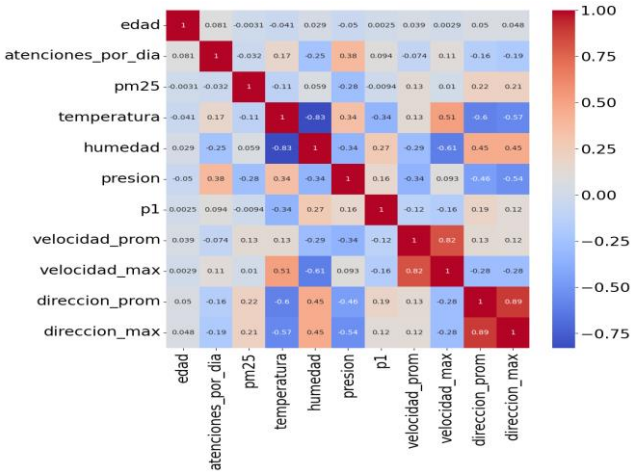

Figura 4 Matriz de correlación de variables seleccionadas.

Seguidamente, estas variables transformadas se llevan a un escalado bajo MinMaxScaler debido a la diferencia de magnitudes de las unidades de algunas variables, finalmente, se dividen los datos en entrenamiento y testeo usando el campo "atenciones_por_dia" como variable de respuesta; esta división de datos se toma un 70% para entrenamiento y un 30% para testeo, además, se deja un valor semilla (random state) para reproducibilidad de 42.

c. Métodos de análisis y Modelado

Los árboles de regresión son un tipo de modelo supervisado que se utiliza para predecir variables numéricas continuas. Funcionan dividiendo el espacio de los datos en regiones más pequeñas gracias al uso de reglas de decisiones, logrando que a cada región se le asigne un valor de predicción basándose principalmente en el promedio de la variable objetivo dentro de esa región (James et al., 2013).

Dado que el objetivo del proyecto se involucra en modelos de regresión al tener una variable objetivo numérica continua, se buscan métodos y modelos correspondientes a esta necesidad. Es por ello que se parte de la intención de usar modelos de árboles que son los más comunes para este tipo de fenómenos. No se consideraron otros tipos de modelos de regresión, como redes neuronales, máquinas de vectores de soporte (SVM) o modelos lineales, debido a que los modelos basados en árboles ofrecen una combinación favorable de rendimiento, interpretabilidad y flexibilidad. A diferencia de las redes neuronales, que requieren una mayor cantidad de datos, preprocesamiento intensivo computacionalmente y cuya estructura resulta difícil de interpretar, otra de las ventajas de usar modelos basados en árboles es que ofrece la posibilidad de manejar predictores tanto numéricos como categóricos, sin necesidad de cumplir una distribución de los datos, identificación de variables importantes, entre otras (Amat, 2020). Además, se debe tener en cuenta que, en el entrenamiento de un árbol de regresión, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta un nodo terminal (ver Figura 5), y la predicción del árbol es la media de la variable de respuesta de las observaciones de entrenamiento de ese nodo terminal. En general, estos modelos de árboles se entrenan en dos etapas: división sucesiva del espacio de los predictores generando regiones no solapantes (nodos terminales) y predicción de la variable respuesta en cada región. La forma de identificar las divisiones del espacio se realiza mediante la suma de cuadrados residuales, ver la ecuación (1).

$$RSS = \sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - \hat{y}_{R_j})^2 \quad (1)$$

De esta ecuación se busca es encontrar las regiones en las que se minimiza la suma de cuadrados residuales, y para ello se usa de la división binaria recursiva donde se busca encontrar en cada iteración el predictor Xj y el punto de corte (umbral) es tal que el sumatorio de las desviaciones al cuadrado entre las observaciones y la media de la región a la que pertenecen sea lo menor posible.

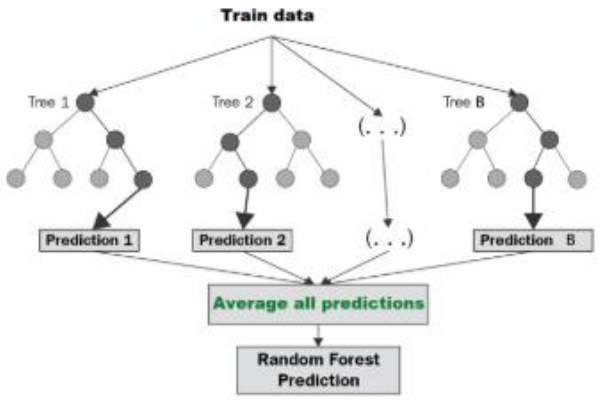

Figura 5 Diagrama general de árboles de decisión (James et al., 2013).

A pesar que los modelos de árboles funcionan de esta misma manera, se encuentran variaciones del modelo que se adaptan a diferentes escenarios, por ejemplo el modelo Random Forest, el cual "está formado por un conjunto (ensemble) de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída de los datos de entrenamiento originales mediante bootstrapping" (Amat, 2020), esto implica que cada árbol se entrena con unos datos ligeramente distintos. Adicionalmente, este modelo permite identificar importancia de las características mediante la librería de scikit-learn en el paquete de RandomForestRegressor que a través del método feature_importances_ establece el grado de importancia de las variables para el modelo a partir de la importancia de Gini (Tiwari, 2023). También, es posible observar la importancia de las variables mediante la técnica de

reducción de dimensionalidad Principal Component Analysis (PCA), la importancia de cada variable viene determinada por la matriz de cargas, esta matriz representa la contribución de cada característica original a los componentes principales. Las características con valores absolutos más altos en las cargas tienen una mayor influencia en el componente principal correspondiente (Jain, 2025).

Respecto a la evaluación del modelo, puesto que es un modelo de regresión se usan métricas como Mean Squared Error (MSE), Root Mean Squared Error (RMSE) y R-squared (R2) en los datos de entrenamiento y testeo para observar presencia de overfitting o underfitting, las ecuaciones objetivo se observan en las ecuaciones (2), (3) y (4).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \qquad (2)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \qquad (3)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \underline{y})^2} \qquad (4)$$

Se considera el uso de estas métricas porque tanto el RMSE como el R-cuadrado cuantifican la precisión con la que un modelo de regresión lineal se ajusta a un conjunto de datos. El RMSE indica la precisión con la que un modelo de regresión puede predecir el valor de una variable de respuesta en términos absolutos, mientras que el R-cuadrado indica la precisión con la que las variables predictoras pueden explicar la variación en dicha variable (Chugh, 2020). Además, estas métricas juegan el papel de selección de modelo tras la comparación entre 3 modelos, XGBoost, Light Gradient y Random Forest.

Para llevar a cabo este proyecto se usa lenguaje Python para el acceso a diferentes librerías de Machine Learning y transformación de datos como scikit-learn y pandas, dichas librerías se suministran bajo un archivo de texto requirements.txt para la instalación del versionamiento de ejecución requerido. A pesar de manejar cantidades de datos del orden más alto de 3 millones, se usó por infraestructura de ejecución una CPU del ambiente de trabajo.

3. Resultados y Discusión

En principio, al llevar a cabo las técnicas de EDA y preprocesamiento para obtener los datos de alimento para el modelado, resultaron 18 campos para implementación (17 características predictoras y 1 variable objetivo). Con estas variables se alimentaron los modelos de XGBoost, Light Gradient Boosting y Random Forest dando como resultados los ajustes de puntos observados en las Figura 6 a Figura 8, respectivamente.

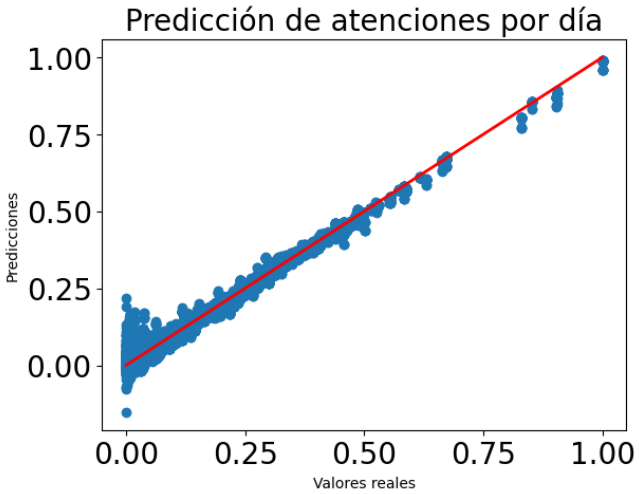

Figura 6 Gráfico de ajuste modelo XGBoost.

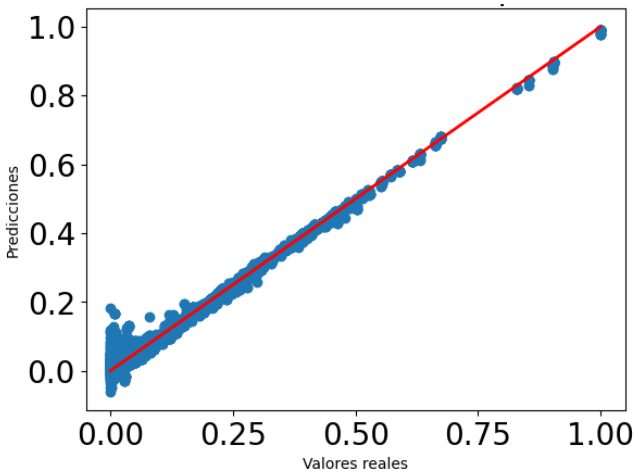

Figura 7 Gráfico de ajuste modelo Light Gradient.

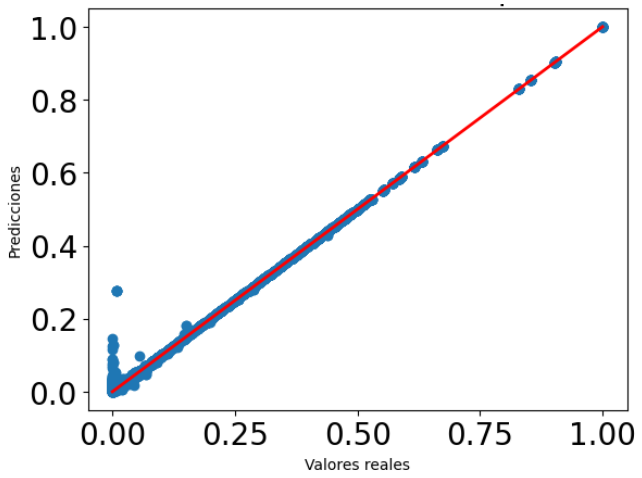

Figura 8 Gráfico de ajuste modelo Random Forest.

Al observar las figuras anteriores se nota una afinidad de ajuste fuerte de los modelos de árboles de regresión a los datos presentados bajo el tratamiento, para compararlos bajo sus métricas se presenta la Tabla 1.

Tabla 1 Cuadro comparativo de métricas de ajuste en modelos.

Modelo	R2	MSE	RMSE	MAE
XGBoost	0.9890	1.95e-04	0.0140	0.0096
Light Gradient	0.9943	1.01e-04	0.0101	0.0069
Random Forest	0.9988	2.13e-05	0.0046	0.0006

Dado que los modelos presentan un ajuste apreciable con el fenómeno y los datos en este tiempo de estudio para la determinación de las atenciones médicas por día, pasamos a una fase de eliminación de variables según la importancia de cada una dentro del modelo, para este fin se utiliza las técnicas de importancia declarada por PCA y por el modelo de Random Forest, obteniendo lo observado en la Tabla 2 y Figura 9.

Tabla 2 Pesos de características por componentes principales PCA.

Características	Peso Variable
NombreIPS	0.999933
estacion	0.006148
comuna	0.006148
sector_ips	0.006148
presion	0.002518
mes	0.002068
dirreccion_prom	0.001906
temperatura	0.001392
humedad	0.001248
velocidad_prom	0.001000
dia_semana	0.000747
pm25	0.000440
dia	0.000238
anio	0.000202
atenciones_por_dia	0.000189
festivo	0.000115
edad	0.000013
p1	0.000012

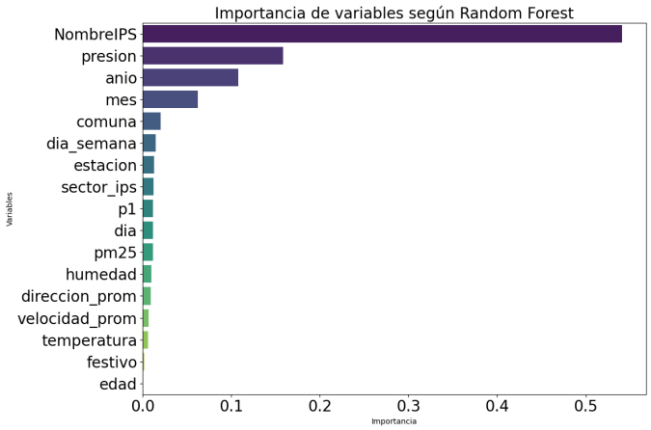

Figura 9 Importancia de las variables según el modelo Random Forest.

Dadas las observaciones presentadas en las figuras anteriores, se realiza un análisis bajo Shap Values con el fin de observar gráficamente la dependencia de las variables y el impacto en las salidas del modelo XGBoost, así se nota la Figura 10, tres representaciones de las variables en dependencia.

Dadas estas observaciones se nota que las 4 variables de mayor impacto en el modelo son “NombreIPS”, “presion”, “anio” y “mes”, por ende, bajo estas visualizaciones se procede a realizar una ejecución de los modelos nuevamente con el uso de estas 4 variables, con el fin de limitar la necesidad de extracción de información para el estimado de la variable de las atenciones médicas diarias. El comparativo de estas métricas bajo la eliminación de las variables se observa en la Tabla 3.

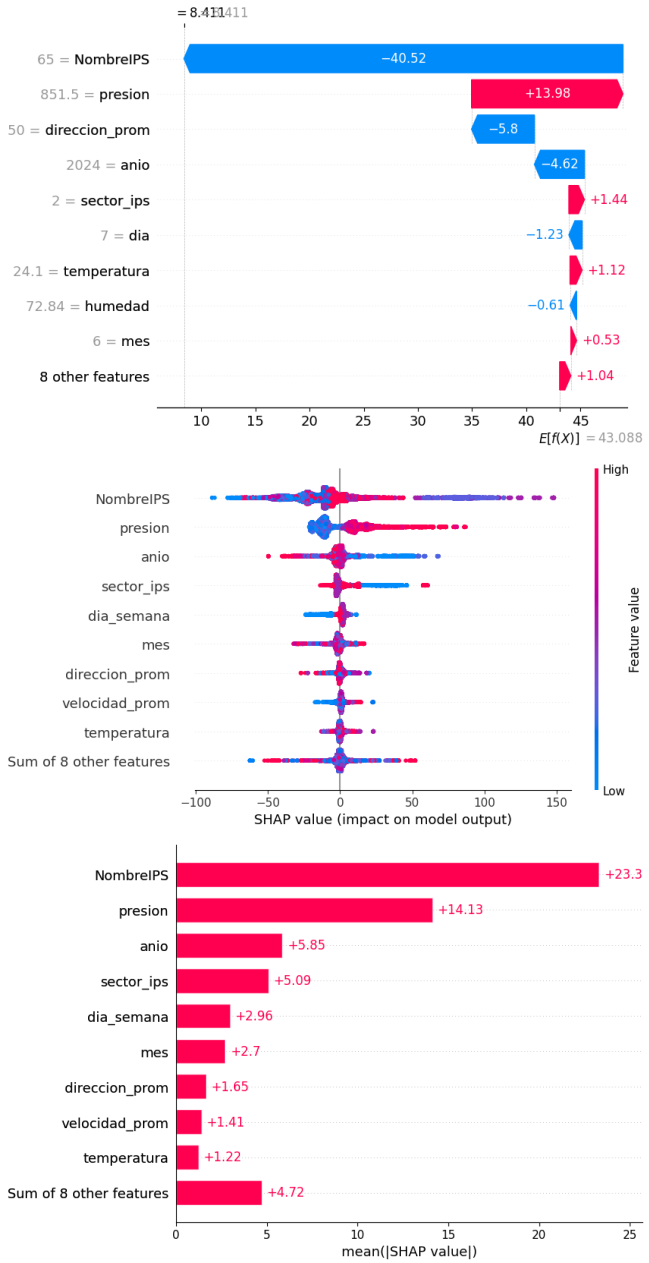

Figura 10 Representaciones del impacto de las variables en el modelo XGBoost por shap values.

Tabla 3 Cuadro comparativo de métricas de ajuste en modelos luego de eliminación de variables de poca importancia.

Modelo	R2	MSE	RMSE	MAE
XGBoost	0.9780	0.0004	0.0198	0.0126
Light Gradient	0.9863	0.0002	0.0156	0.0097
Random Forest	0.9940	0.0001	0.0103	0.0020

Con estos hallazgos se establece que el modelo de Random Forest es aquel de mejor ajuste sobre los datos luego de la eliminación de las variables.

Dados los resultados anteriores podemos notar que el ingreso de atenciones médicas por diagnósticos asociados a enfermedades respiratorias se puede ver influenciado por el sitio de consulta, los efectos en el incremento de la presión (al estar correlacionada con la variable objetivo) y la época del año. Esto bajo los resultados obtenidos de la Figura 10. donde se detalla un shap value de 23.3 sobre la variable 'NombreIPS', asignándole como la variable de mayor varianza así como lo confirmado en la Tabla 2 Figura 9; seguido de ello, se tiene la variable 'presion' con un impacto promedio sobre la predicción de 14.13 que a altos valores de la característica representa un aumento de la predicción y en sentido contrario, una presión baja disminuye directamente la predicción. Por otro lado, el tema temporal dado por las variables 'anio' y 'mes' implican que a valores de año más altos generan shap values positivos, indicando que en años más recientes el número de atenciones previstas tiende a subir.

Ya que el modelo de Random Forest es aquel que presenta mayor métrica de ajuste de 0.9940 podemos interpretar que un modelo de bagging (bootstrap + aggregating) puede ajustarse en mejor proporción a los otros contemplados dado que al entrenar todos los árboles en paralelo sobre muestras y subconjuntos de features diferentes, el modelo puede tender a ser más robusto frente al ruido extremo, lo que hace que generalice en mejor medida por un efecto "de serie", adicional, dado que solo se tiene 4 variables Random Forest permite realizar una mayor distribución de las variables en los árboles y evita que una variable domine todos los árboles. Además, el modelo usado se entrenó bajo hiperparámetros predeterminados, lo cual no requiere usar técnicas de optimización.

Es importante notar que en el proceso de importancia de las variables se eliminaron 12 campos de análisis que permiten un ajuste mayor según la Tabla 1, dichos campos hacen relación a condiciones ambientales y locativas que podrían ampliar la explicación fenomenológica del modelo, pero que por facilidades de adquisición de información y manipulación del modelo al no verse una afectada en mayor proporción de la precisión se decide recurrir a este mínimo de variables. Bajo posteriores estudios se espera ampliar el reconocimiento espacial de estaciones de medición para abarcar mayores áreas de condiciones y usar datos que

quedaron restantes de lugares de atención hospitalaria por falta de asignación de estación de monitoreo, además, estudiar la posibilidad del uso de valores con calidad de dato (proporcionado por el SIATA) de mayor clasificación con el fin de obtener mayor cantidad de datos de un tiempo específico. Con estos datos que abarcan mayor extensión del fenómenos en las diferentes áreas se podría realizar un agrupamiento mediante técnicas de aprendizaje no supervisado con el fin de identificar zonas similares en términos de condiciones ambientales y de salud, de tal forma, que se permita tener un monitoreo espacial de zonas de mayor contaminación que impliquen ingresos superiores en las atenciones médicas, llegando a un análisis causal de incrementos espontáneos por las diferentes condiciones, encontrando atípicos en esta serie de tiempo.

4. Conclusiones

Es importante notar que los campos que se mantienen luego de la eliminación de características de baja importancia son variables que hacen sentido a una condición meteorológica, una observación de localización como la IPS de consulta y de una especificidad temporal bajo el año y mes de toma de dato, lo que hace sentido físico dado el acontecimiento climático que influye en la cantidad de pacientes que ingresan a las instituciones de salud.

El modelo Random Forest es aquel que tiene mayor ajuste sobre las variables de importancia en comparación a los otros propuestos en este trabajo, bajo una precisión de R2 de 0.9940 el modelo permite pronosticar con un día de antelación la cantidad de atenciones médicas por diagnósticos relacionados con enfermedades respiratorias respecto a la IPS de mayor cercanía. Además, el modelo muestra menor dispersión de los datos respecto al ajuste del modelo en entrenamiento y testeo, por ende, se toma como el modelo de mayor ajuste sobre los datos utilizados.

Como opción posterior se podría pensar en un despliegue de consulta bajo entidades de monitoreo de condiciones climáticas y establecimientos de salud, de tal forma que se realice un alertamiento con antelación que permita actuar de forma rápida y eficiente sobre aquellas IPS que requieran mayor capacidad, además, se podría tomar como una herramienta para el cálculo de presupuestos en diferentes tipos de pacientes, distribuyendo de forma más adecuada recursos para diferentes áreas de las entidades.

5. Bibliografía

Área Metropolitana Valle de Aburrá. (n.d.). ¿Qué es SIATA? Área Metropolitana SIATA. Retrieved November 20, 2024, from https://www.metropol.gov.co/ambiental/siata/paginas/que-es.aspx

Hospital internacional de Colombia. (2024, Agosto 26). Contaminación del aire y enfermedades respiratorias: ¿cómo afecta a tu salud? Neumología y Salud Respiratoria. Retrieved Noviembre 16, 2024, from https://hic.fcv.org/co/blog/neumologia-y-salud-respiratoria/contaminacion-del-aire-y-enfermedades-respiratorias-como-afecta-a-tu-salud#:~:text=Adem%C3%A1s%2C%20la%20contaminaci%C3%B3n%20del%20aire,aumentar%20el%20riesgo%20de%20infecciones

Ministerio de Salud y Protección Social. (2022, November 17). Muertes por enfermedades respiratorias crónicas han disminuido. Ministerio de Salud y Protección Social. Retrieved November 16, 2024, from https://www.minsalud.gov.co/Paginas/Muertes-por-enfermedades-respiratorias-cronicas-han-disminuido.aspx

Ministerio de Salud y Protección Social - República de Colombia. (n.d.). Sistema de Información de Prestaciones de Salud - RIPS. Ministerio de Salud y Protección Social. Retrieved November 20, 2024, from https://www.minsalud.gov.co/proteccionsocial/paginas/rips.aspx

Agencia de Protección Ambiental de Estados Unidos. (2024, June 25). Conceptos básicos sobre el material particulado (PM, por sus siglas en inglés) | US EPA. EPA en español. Retrieved September 5, 2024, from https://espanol.epa.gov/espanol/conceptos-basicos-sobre-el-material-particulado-pm-por-sus-siglas-en-ingles

Área Metropolitana Valle del Aburrá. (2019). Calidad del aire. Área Metropolitana. Retrieved September 7, 2024, from https://www.metropol.gov.co/ambiental/calidad-del-aire

Baena, D., Jiménez, J., Zapata, C., & Ramirez, Á. (2019, Marzo 25). Red neuronal artificial aplicado para el pronóstico de eventos críticos de PM2.5 en el Valle de Aburrá. SciELO Colombia. Retrieved September 10, 2024, from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0012-73532019000200347#t1

BBC News Mundo. (2017, November 15). Qué es EPOC, la enfermedad pulmonar silenciosa que mata a 3

millones de personas al año y no tiene cura. BBC. Retrieved September 7, 2024, from https://www.bbc.com/mundo/noticias-41997332

Ministerio de Salud y Protección Social. (2022, November 17). Muertes por enfermedades respiratorias crónicas han disminuido. Ministerio de Salud y Protección Social. Retrieved September 6, 2024, from https://www.minsalud.gov.co/Paginas/Muertes-por-enfermedades-respiratorias-cronicas-han-disminuido.aspx

Minsalud. (2021, August 23). Minsalud comprometido con la calidad del aire. Ministerio de Salud y Protección Social. Retrieved September 9, 2024, from https://www.minsalud.gov.co/Paginas/Minsalud-comprometido-con-la-calidad-del-aire-.aspx

Parra, J., Oviedo, A., & Omaya, F. (2020, Octubre 25). Analítica de datos: incidencia de la contaminación ambiental en la salud pública en Medellín (Colombia). Rev. Salud Pública, 22(6), 609-617. https://doi.org/10.15446/rsap.V22n6.78985

Pérez, J., Montoya, P., Sanchez, J. M., Hernández, S., & Ramírez, M. (2023, Diciembre 14). Forecasting 24 h averaged PM2.5 concentration in the Aburrá Valley using tree-based machine learning models, global forecasts, and satellite information. ASCMO - Volumes. Retrieved September 10, 2024, from https://ascmo.copernicus.org/articles/9/121/2023/ascmo-9-121-2023.html

Torres Jaramillo, J. A. (2021, September 6). El aire nuestro de cada día. Instituto de Ciencias de la Atmósfera y Cambio Climático. Retrieved September 5, 2024, from https://www.atmosfera.unam.mx/el-aire-nuestro-de-cada-dia/

Amat, J. (2020, Octubre). Arboles de decision, Random Forest, Gradient Boosting y C5.0. cienciadedatos.net. Retrieved May 27, 2025, from https://cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting

Chugh, A. (2020, December 7). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? Medium. Retrieved May 27, 2025, from https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Jain, S. (2025, May 18). Feature Importance in PCA: Analyzing Loadings and Biplots. GeeksforGeeks. Retrieved May 27, 2025, from https://www.geeksforgeeks.org/feature-importance-in-pca-analyzing-loadings-and-biplots/

Tiwari, R. (2023, April 18). Using Random Forest For Feature Importance And Feature Selection. Medium. Retrieved May 27, 2025, from https://medium.com/@prasannarghattikar/using-random-forest-for-feature-importance-118462c40

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer., tomado de https://www.statlearning.com